JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 3/4, 2000
Mary Ann Liebert, Inc.
Pp. 469–481

Matching Simulation and Experiment: A New Simplified Model for Simulating Protein Folding

JON M. SORENSON[1] and TERESA HEAD-GORDON[2]

ABSTRACT

Simulations of simplified protein folding models have provided much insight into solving the protein folding problem. We propose here a new off-lattice bead model, capable of simulating several different fold classes of small proteins. We present the sequence for an α/β protein resembling the IgG-binding proteins L and G. The thermodynamics of the folding process for this model are characterized using the multiple multihistogram method combined with constant-temperature Langevin simulations. The folding is shown to be highly cooperative, with chain collapse nearly accompanying folding. Two parallel folding pathways are shown to exist on the folding free energy landscape. One pathway contains an intermediate—similar to experiments on protein G, and one pathway contains no intermediates—similar to experiments on protein L. The folding kinetics are characterized by tabulating mean-first passage times, and we show that the onset of glasslike kinetics occurs at much lower temperatures than the folding temperature. This model is expected to be useful in many future contexts: investigating questions of the role of local versus nonlocal interactions in various fold classes, addressing the effect of sequence mutations affecting secondary structure propensities, and providing a computationally feasible model for studying the role of solvation forces in protein folding.

Key words: off-lattice models, protein L, protein G, protein folding, multiple histogram method, multi-state kinetics.

INTRODUCTION

UNDERSTANDING HOW AN UNFOLDED POLYPEPTIDE CHAIN folds in solution quickly and correctly to form the proper tertiary structure is a long-standing problem in biophysical chemistry. In recent years, considerable insight into this process has stemmed from experimental folding studies on small proteins (Fersht, 1997; Eaton *et al*., 1998; Capaldi and Radford, 1998) combined with new advances in theoretical perspectives (Dill and Chan, 1997; Onuchic *et al*., 1997) and computational approaches (Dill *et al*., 1995; Shakhnovich, 1997; Dobson *et al*., 1998).

[1]Department of Chemistry, University of California, Berkeley, CA 94720.
[2]Physical Biosciences and Life Sciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

As computer speed increases and molecular force fields become increasingly more sophisticated, simulations of the folding process promise to answer many of the outstanding questions left in understanding the protein folding problem. However, atomistic simulations of proteins in explicit solvent, while now performable (Duan and Kollman, 1998), are still computationally extremely demanding. This issue is heightened when we consider that extracting reproducible conclusions from folding simulations requires studying many trajectories, possibly in a variety of conditions. In order to make useful generalizations and collect meaningful statistics from our computational models, we still require simpler, computationally faster models. In addition, a central tenet of theoretical research is the desire to extract the underlying principles of a physical mechanism in its simplest form. We wish to construct models which capture the essential physics of the protein folding problem while maintaining a simplicity which allows clean analysis and provides easily generalizable insight (Shakhnovich, 1996). In addition, we desire the retention of enough chemical detail to allow direct comparison to experiment for general trends in thermodynamics and kinetics.

In this paper we present a new simplified off-lattice model for studying protein folding. The model is based on a previous useful model developed by Thirumalai and coworkers (Honeycutt and Thirumalai, 1990; Guo and Thirumalai, 1994) for studying all-β proteins or all-α (Guo and Thirumalai, 1996) proteins. Our model is intended as a hybrid of the previous all-β and all-α models, and as such is capable of simulating small all-β, all-α, and mixed α/β proteins. Encompassing these major fold classes under one framework facilitates comparison between the folding of different topologies and allows greater flexibility in comparison to experiment.

As an example of the power of this new model, we designed a sequence which folds to a structure with the same overall topology of the IgG-binding proteins L and G (Ramírez-Alvarado *et al*., 1997; Kim *et al*., 1998a). These two small proteins, while having little sequence homology, have nearly identical structures, consisting of a central α-helix packed against a mixed β-sheet composed of two β-hairpin structures. Both proteins make excellent targets for theoretical study, as the folding of both protein L and protein G has been extensively studied by experiment, with many useful mutation studies (Gu *et al*., 1997; Kim *et al*., 1998b; Kim *et al*., 1998a), secondary structure fragment studies (Blanco and Serrano, 1995; Ramírez-Alvarado *et al*., 1997; Blanco *et al*., 1997), and other biophysical investigations (Park *et al*., 1997; Plaxco *et al*. 1999).

Our simplified model enables us to analyze many trajectories, to collect good statistics for fully characterizing the thermodynamics and kinetics of folding, and to quickly explore modifications to the original sequence. In addition, the model possesses sufficient complexity to investigate questions about the role of turn propensities in small α/β proteins (Gu *et al*., 1997) and the effect of sequence mutations which change the distribution of nonlocal contacts or destabilize secondary structure (Kim *et al*., 1998b).

In this paper, we first present the model and our methods for simulating the folding process and collecting thermodynamic and kinetic information. We next present the sequence which folds to a protein L/G-like structure and characterize the thermodynamics and kinetics of this process. Two competing pathways—one with an intermediate and one without—are shown to contribute to the folding process. We conclude with a summary of these results and a look at future directions with this new off-lattice model.

MODEL AND METHODS

The Hamiltonian

The energy function for the model is proposed in the spirit of the off-lattice bead models developed by Thirumalai and coworkers Honeycutt and Thirumalai, 1990; Guo and Thirumalai, 1994; Guo and Thirumalai, 1996), and as such we took most of the parameters for the Hamiltonian from their previous work to allow relevant comparison. A similarly spirited extrapolation from their work has also been proposed by Ferguson and Garrett (1999).

The protein chain is modeled as a chain of beads of three flavors: hydrophobic (B), hydrophilic (L), or neutral (N). Attraction between the hydrophobic beads provides the energetic driving force for formation of a strong core, repulsion between the hydrophilic beads and other beads are used to balance the forces and bias the correct native fold, and the neutral beads serve as soft spheres with little repulsion and typically signal the turn regions in the sequence.

The Hamiltonian for the model is

$$H = \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} \{A[1 + \cos\phi] + B[1 - \cos\phi] + C[1 + \cos\ 3\phi] + D[1 + \cos(\phi + \pi/4)]\} + \sum_{i,j \geq i+3} 4\epsilon_H S_1 \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - S_2 \left(\frac{\sigma}{r_{ij}} \right)^{6} \right]. \quad (1)$$

Bond lengths are held rigid, and the bond angles are maintained by a harmonic potential with force constant $k_\theta = 20\epsilon_H/(\text{rad})^2$ and equilibrium bond angle $\theta_0 = 105°$.

The dihedral potential is a combination of the potentials used in the previous all-β (Guo and Thirumalai, 1994) and all-α (Guo and Thirumalai, 1996) models. Each dihedral in the chain is predefined to be either helical ($A = 0, B = C = D = 1.2\epsilon_H$), extended ($A = 0.9\epsilon_H, C = 1.2\epsilon_H, B = D = 0$), or turn ($A = B = D = 0, C = 0.2\epsilon_H$). These potentials correspond closely to the potentials employed in the previous studies; the most notable difference is that the β-sheet dihedral potential uses a slightly different value for A. This new parameter choice was set to make the helical and extended dihedral potentials more comparable to each other in terms of minima stability and barrier heights.

The nonlocal interactions are given by $S_1 = S_2 = 1$ for BB interactions, $S_1 = 1/3$ and $S_2 = -1$ for LL and LB interactions, and $S_1 = 1$ and $S_2 = 0$ for all interactions involving N residues. These interactions are also interpolations between the nonlocal interactions in the all-β (Guo and Thirumalai, 1994) and all-α (Guo and Thirumalai, 1996) models. Unlike with the all-α model Guo and Thirumalai, 1996), no new scaling factor is introduced to balance the nonlocal interactions versus the dihedral biases.

The chief difference in this new formulation is the assignment of dihedral potentials. In the previous models, the turn dihedrals in the turn regions were signaled by the presence of neutral beads in the primary sequence. All other dihedrals were assumed to be of the type for the model being studied (either all-β or all-α). To study mixed α/β proteins, we need to separate the specification of dihedral potentials from the primary sequence. In addition to specifying a primary sequence, a model sequence is also defined by specifying a sequence of secondary structure propensities. Since the model lacks important determinants of protein structure, such as side-chain packing and backbone hydrogen bonding, the dihedral potentials serve as potentials of mean force, enforcing the secondary structure propensity that would be there if side chains and/or the ability to form backbone hydrogen bonds were present. In particular, the formation of β-sheets is driven by the attraction of B beads and the action of the extended dihedral potentials. In this sense, the B beads represent generic attractive forces responsible for β-sheets in real proteins such as hydrophobic forces and hydrogen bond formation and the dihedral potential represents the intrinsic propensity for some amino acid sequences to form extended structures.

There are several advantages to this formulation. By separating nonlocal (bead–bead) and local (dihedral) interactions, we can easily vary the relative strengths of these interactions and investigate questions of the role of local interactions versus nonlocal interactions (Chou and Shakhnovich, 1999). While this has not been performed in this current study, a thorough examination of the effects of parametrization of this model could address this point as well as highlight which features of the model are artifacts of the parametrization and which are to be taken as more robust. Another benefit of separating the dihedral biases is that, by allowing turn regions to contain not only turn dihedrals, we can examine the effect of changing secondary structure propensity in turn regions. This is similar to the use of a helical dihedral introduced into the the turn region of the all-α model (Guo and Thirumalai, 1996). Similarly we can examine the effect of destabilizing secondary structure elements by introducing turn dihedrals into helical or sheet regions. Thus the separation of dihedral biases from the primary sequence can mimic mutation studies which change the local secondary structure propensity of a protein sequence (Gu *et al.*, 1997; Kim *et al.*, 1998b). This separation is also responsible for increasing the complexity of sequences in the model, allowing, in effect, more than the three flavors of beads governing nonlocal interactions. The combinatorial possibilities of changing bead type plus dihedral bias makes the model more like a bead model with a larger number of flavors with the expected benefits of less native-state degeneracy (Shakhnovich, 1994; Yue *et al.*, 1995; Wolynes, 1997) and larger energy gaps between native and nonnative states (Shakhnovich, 1997; Sorenson and Head-Gordon, 1998).

Simulation methods

We use a simulated annealing protocol to find the global minimum for sequences in this model. Once a global minimum is found, constant-temperature Langevin simulations are carried out for characterizing the thermodynamics and kinetics of folding to the native state. The protocol for both the simulated annealing and Langevin simulations has been described extensively in a previous publication (Sorenson and Head-Gordon, 1999). The simulations are performed in reduced units, with the units of mass m, length σ, energy ϵ_H, and k_B all set equal to one; temperature is in units of ϵ_H/k_B. The unit of reduced time is $\tau = \sqrt{m\sigma^2/\epsilon_H}$.

The free energy landscape of the model is characterized by sampling using the multiple multidimensional histogram method (Ferrenberg and Swendsen, 1989; Kumar *et al*., 1995). The use of this method to extract the density of states for protein folding systems has been well documented (Socci and Onuchic, 1995; Guo and Brooks, 1997; Sorenson and Head-Gordon, 1999). In this work, we collected six-dimensional histograms over energy and five order parameters—radius of gyration (R_g), χ, χ_H, χ_{β_1}, and χ_{β_2}. χ is the order parameter for folding to the native state (Guo and Thirumalai, 1994):

$$\chi = \frac{1}{M} \sum_{i,j \geq i+4}^{N} \theta\left(\epsilon - |r_{ij} - r_{ij}^{nat}|\right) \tag{2}$$

where the double sum is over beads on the chain, r_{ij} and r_{ij}^{nat} are the distances between beads i and j in the state for comparison and the native state, respectively, θ is the Heaviside step function, and $\epsilon = 0.2$ to account for small fluctuations away from the native-state structure. M is a normalizing factor to ensure that $\chi = 1$ when the chain is identical to the native state and $\chi \approx 0$ when the chain is in a random coil state. The other three χ's are a specialization of χ to monitor formation of specific secondary structure elements. The sum over residues for χ_H involves only beads in the helix, and the sums for χ_{β_1} and χ_{β_2} involve only beads in β-hairpins one and two, respectively.

The kinetics of the folding process can be characterized by calculating first passage times. These times were calculated by taking high temperature unfolded structures and recording the time that they first folded to the native state at a given temperature. The native state is defined in this study as $\chi \geq 0.42$, as this value for χ corresponds to the dividing surface between the nonnative and native basins of attraction. This value is determined from plotting free energy as a function of native-state similarity (Guo and Brooks, 1997; Sorenson and Head-Gordon, 1999). First passage times for chain collapse ($R_g^2 \leq 8.0\sigma^2$) were also collected to investigate questions of how closely chain collapse accompanies folding to the native state (Plaxco *et al*., 1999).

THE MIXED α/β MODEL

The sequence and secondary structure propensities for the the mixed α/β model are presented in Table 1. The corresponding global minimum structure found from simulated annealing is shown in Figure 1. Also shown for comparison is the NMR-derived solution structure of protein L (Wikstrom *et al*., 1994) (omitting the first seventeen residues of the disordered N-terminus). The overall topologies of the two structures are very similar, both consisting of a central α-helix packed against a mixed β-sheet formed by two β-hairpin structures. The main difference is that, in the model, the two β-hairpins pack together to maximize the

Table 1. Primary Sequence and Dihedral Biases for the Mixed α/β Model. The Secondary Structure Elements Correspond to s = sheet, h = helix, and t = turn

Sequence
LBLBLBLBBNNNBBBLBBBBBNNNLLBLLBBLLBNBLBLBLBBNNNBBBLBLBIBL
Secondary structure
sssssssstthsssssssshhshhhhhhhhhhshtssssssshttsssssssss

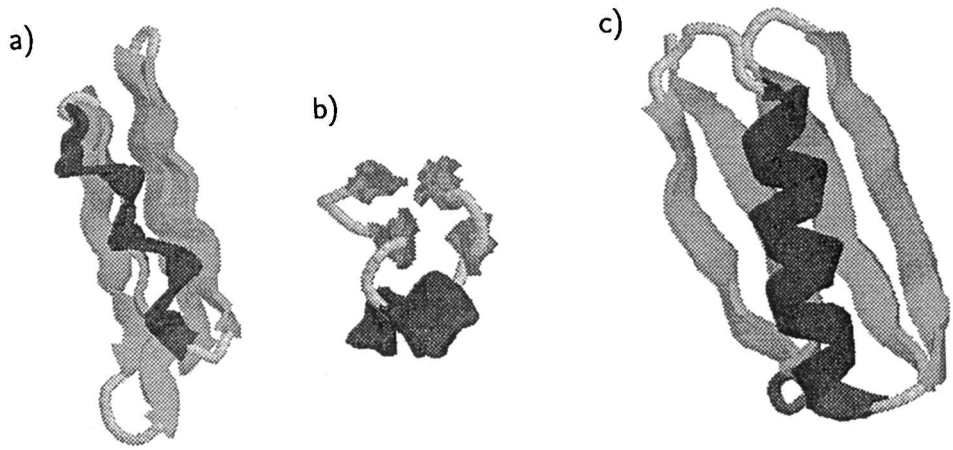

FIG. 1. **(a)** Side view of the native state structure for the mixed α/β model. **(b)** Top view of the same native state. **(c)** NMR solution structure of protein L. The model and protein L are shown in the same representation to emphasize their similar arrangement of secondary structure.

interactions of the hydrophobic core formed between them. In the protein L structure, the sheets do not pack like this; this is most likely due to the bulky steric constraints of the sidechains comprising the hydrophobic core, not present in our bead model.

Thermodynamics

The thermodynamics of the folding process were characterized by simulations over a range of temperatures. Sampling at various temperatures was combined into one picture by use of the multiple histogram method (Ferrenberg and Swendsen, 1989).

Figure 2 shows the heat capacity (C_V) versus temperature for this model. Two very distinct peaks are present, indicating the presence of two transitions as temperature is lowered. We can identify the first

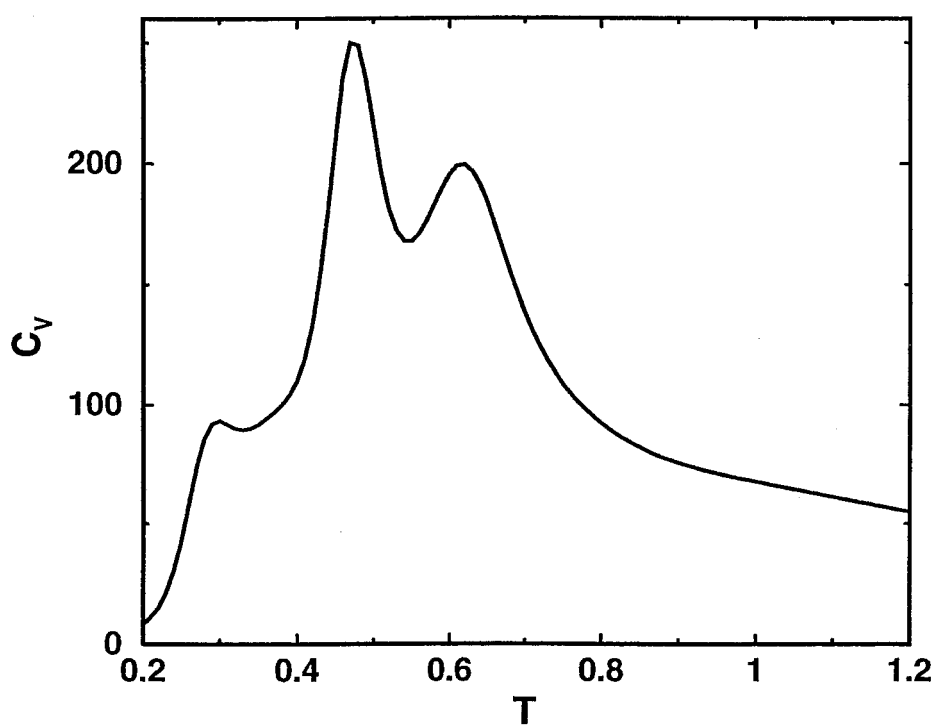

FIG. 2. Heat capacity versus temperature for the mixed α/β model.

transition with the early formation of helical secondary structure. This can be seen in Figure 3 where the values of $\langle \chi_H \rangle$, $\langle \chi_{\beta_1} \rangle$, and $\langle \chi_{\beta_2} \rangle$ versus temperature are plotted. Around $T = 0.62$ we see a weak transition with the formation of some helical structure, but without a corresponding formation of the beta hairpin structures.

As the temperature is further lowered, we approach the major peak in the heat capacity curve at $T = 0.46$. Now all three secondary structure order parameters show a sharp transition and the formation of nativelike structure. Further evidence that $T = 0.46$ is the folding temperature (T_f) follows from the observation that the fluctuations in χ versus temperature ($(\Delta\chi)^2 = \langle \chi^2 \rangle - \langle \chi \rangle^2$) have a sharp and narrow peak at $T = 0.46$ (Sorenson and Head-Gordon, unpublished).

It is notable that the major peak in the heat capacity curve corresponds to the folding transition. This is in sharp contrast to folding in the previous all-β models (Sorenson and Head-Gordon, 1999; Guo and Brooks, 1997), where the major peak in the heat capacity curve corresponds to the collapse transition and the folding transition has an extremely weak heat capacity peak at best. This coincidence of a maximum in heat capacity with the folding transition is strong support for the high cooperativity of the folding process in this model. The collapse transition in this model is at an only slightly higher temperature than the folding transition. The collapse transition temperature is obtained from observing that the fluctuations in the radius of gyration as a function of temperature have a peak at $T_\theta = 0.5$ (Sorenson and Head-Gordon, unpublished), very close to the folding temperature.

We can also see from Figure 3 that helix formation is a less cooperative process in this model than the β-hairpin formation. The final transition to native helix structure is less sharp, and the final amount of native helix content remains relatively low. This could be expected from the model in that helix formation is entirely driven by independent dihedral potentials; no extra helix cooperativity terms, such as $i, i+3$ or $i, i+4$ terms (Kolinski *et al.*, 1996; Liwo *et al.*, 1998), or stabilizing side-chain interactions are present. In contrast, the β-hairpins have the opportunity to fold cooperatively by forming contacts first at the top of the loop, reducing chain entropy and allowing the chain to quickly "zip" up the remaining contacts (Muñoz *et al.*, 1997; Eaton *et al.*, 1998).

At temperatures below the folding transition, it appears that the content of native secondary structure reaches a plateau. This is likely an artifact of the sampling procedure. In order to apply the histogram method, we must simulate at temperatures near the folding temperature or higher; at these relatively high temperatures, the global minimum native state structure is very rarely sampled. Better sampling of low

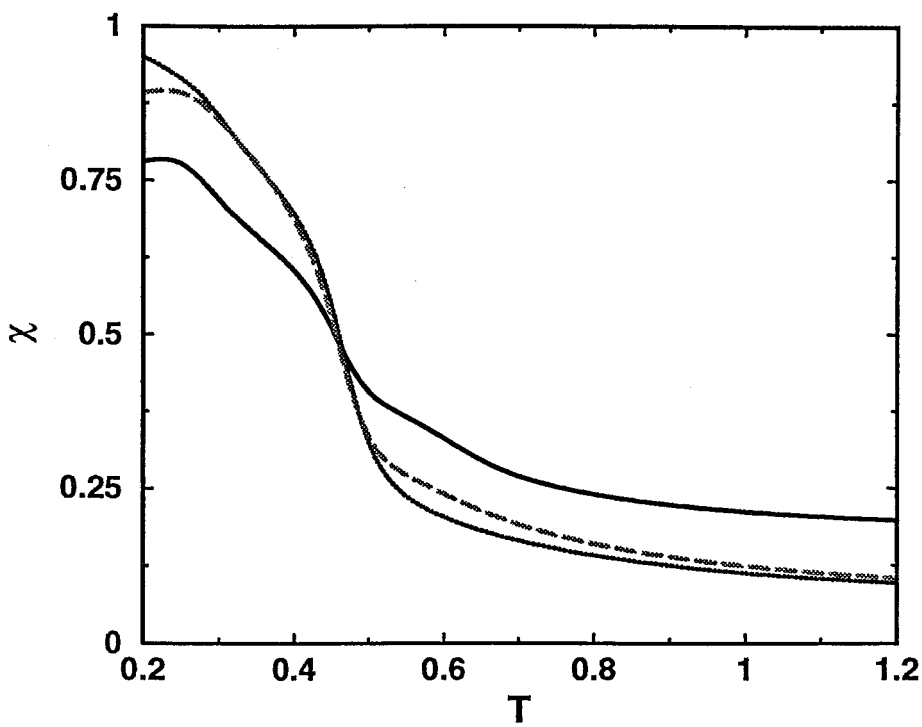

FIG. 3. Formation of secondary structure versus temperature: $\langle \chi_H \rangle$ (solid), $\langle \chi_{\beta_1} \rangle$ (dotted), $\langle \chi_{\beta_2} \rangle$ (dashed).

temperature states would sample these structures and the curves in Figure 3 would converge to 1 at low temperatures. This issue is discussed in greater detail in a previous publication (Sorenson and Head-Gordon, 1999).

Using the multiple multidimensional histogram method (Kumar *et al*., 1995), we can project the underlying free energy landscape onto various combinations of the order parameters. Previous work has found it useful to investigate the potential of mean force as a function of R_g and χ (Guo and Brooks, 1997; Sorenson and Head-Gordon, 1999; Shea *et al*., 1998). The radius of gyration monitors the overall compactness of the chain, while χ measures the similarity of the chain configuration to the native state. If compaction of the chain is always accompanied by formation of native structure, we would expect a diagonal free energy surface. Regions off of the diagonal in the free energy plot show potential traps where compact nonnative states are accessible. The folding free energy landscape for this model is shown in Figure 4. In contrast to the all-β model (Guo and Brooks, 1997; Sorenson and Head-Gordon, 1999), the free energy surface shows more accessible states closer to the diagonal. In particular, states exist with relatively noncompact structure, but partly native structure (in the $R_g \approx 4.25\sigma$, $\chi \approx 0.5$ region). These states would be expected to play a large role in a collapse-concomitant-with-folding scenario. Also, compared to the all-β model, less misfolded traps exist with very compact but nonnative structure.

Since our histograms are sampled over five order parameters characterizing the folding of the chain, we have many possible projections of the free energy landscape onto these parameters. An interesting projection is the potential of mean force as a function of β-hairpins #1 and #2 formation (Figure 5). We can see from the projection that the underlying free energy surface is not symmetrical with respect to β-hairpin formation. It is decidedly more favorable to fold β-hairpin #1 first and then form β-hairpin #2. A similar asymmetry in β-hairpin formation has been previously noted in experiments on protein L (Guo *et al*., 1997).

Examining folding trajectories, we have seen that many folding events follow a pathway whereby β-hairpin #1 forms concurrently with α-helix formation and then β-hairpin #2 folds to form the native

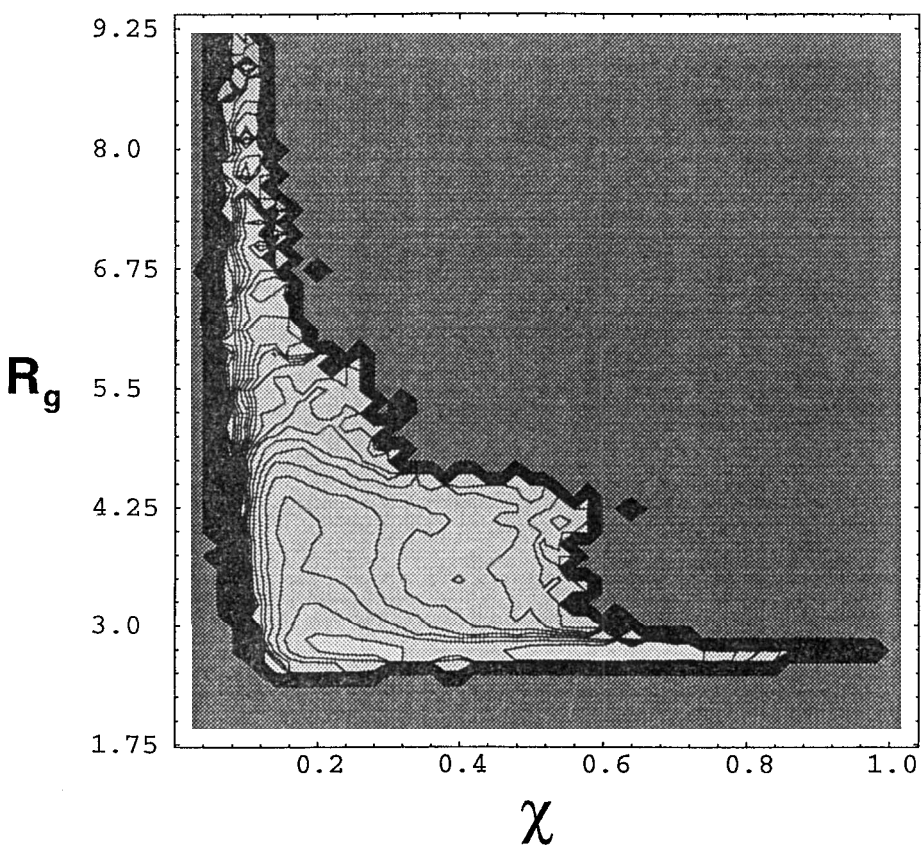

FIG. 4. Free energy at $T = T_f$ as a function of R_g and χ for the mixed α/β model. The contour lines are spaced at intervals of $k_B T$.

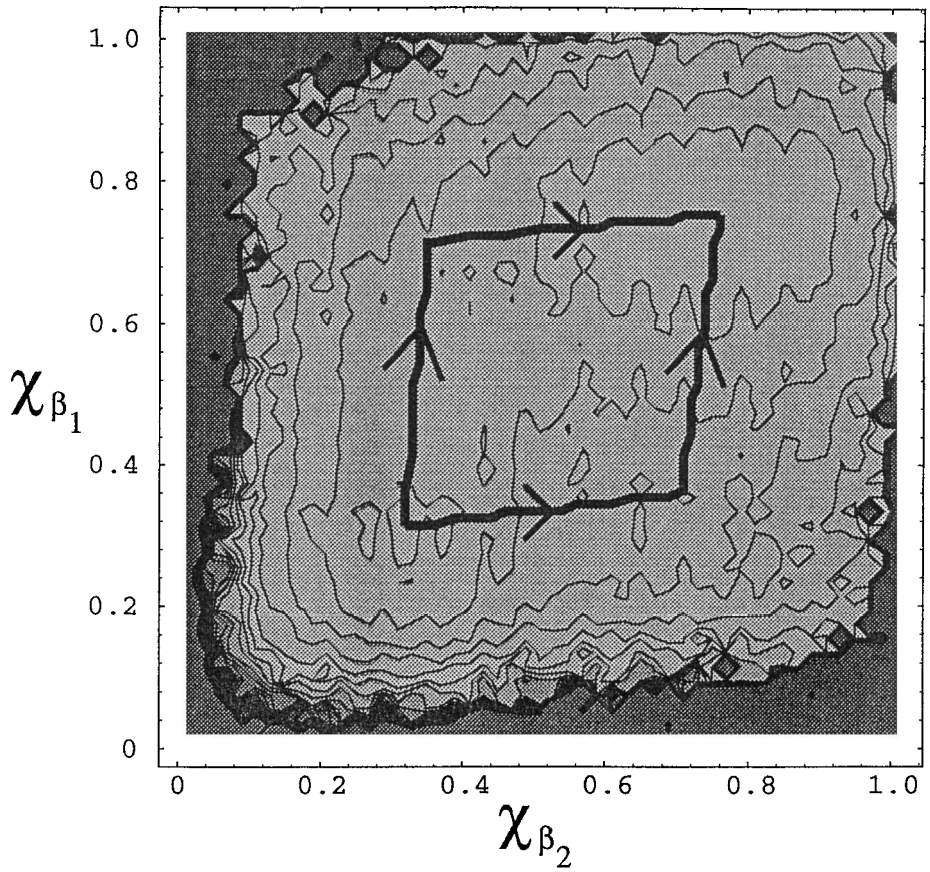

FIG. 5. Free energy at $T = 0.43$ as a function of χ_{β_1} and χ_{β_2} for the mixed α/β model. The contour lines are spaced at intervals of $k_B T$. The upper path is folding path number one, and the lower path is folding path number two.

state. The structure with β-hairpin #1 and the α-helix formed is relatively long-lived and constitutes an intermediate state on the folding pathway.

However, we also have observed folding trajectories which form β-hairpin #2 and the α-helix first and then fold β-hairpin #1 to form the native state. These trajectories quickly cross from the unfolded state to the folded state, with no evidence for a stable β-hairpin #2/α-helix intermediate. This is consistent with the underlying free energy landscape seen in Figure 5.

Assembling these observations, we find that at least two distinct pathways exist for the folding of the mixed α/β structure. One pathway involves formation of an intermediate consisting of β-hairpin #1 and the α-helix. The other pathway involves a single barrier and proceeds by formation of the other β-hairpin structure first. We can quantify these pathways by determining the potential of mean force for two hypothetical pathways illustrating these scenarios (depicted in Figure 5). Given a pathway in $(\chi_{\beta_1}, \chi_{\beta_2})$-space, the potential of mean force is given by

$$w(\chi^i_{\beta_1}, \chi^i_{\beta_2}) = -\frac{1}{\beta} \ln \left[\int dE dR_g d\chi d\chi_H d\chi_{\beta_1} d\chi_{\beta_2}\, \delta(\chi_{\beta_1} - \chi^i_{\beta_1}) \delta(\chi_{\beta_2} - \chi^i_{\beta_2}) \Omega(E, R_g, \chi, \chi_H, \chi_{\beta_1}, \chi_{\beta_2}) e^{-\beta E} \right] \quad (3)$$

where the set $(\chi^i_{\beta_1}, \chi^i_{\beta_2})$ defines the pathway, β is the inverse temperature, and the δ-functions in the integral are approximated as Gaussian functions. Approximating the δ-functions is necessary since the density of states data is tabulated on a grid and the continuous integrals implied by Equation 3 are not possible. Performing this calculation for the two pathways we find the free energy curves shown in Figure 6 at $T = 0.43$, just below the folding temperature. The Gaussian width used for these curves was $\sigma = 0.04$, although similar curves are obtained for varying values of σ from 0.02 to 0.05. As expected, pathway

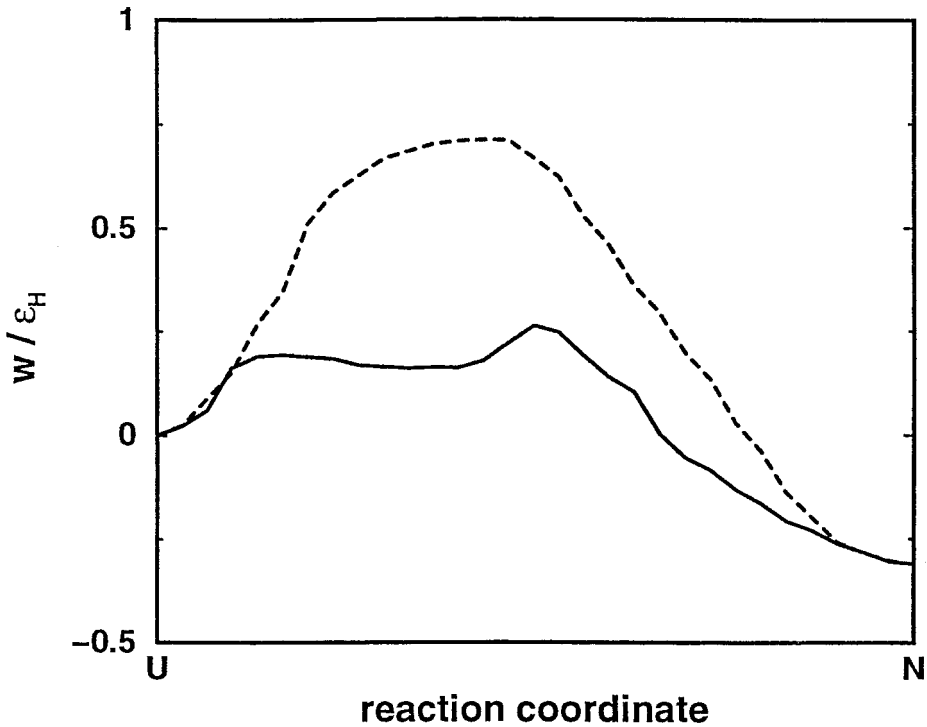

FIG. 6. Potential of mean force at $T = 0.43$ for pathways one (solid line) and two (dashed line) depicted in Figure 5.

one has an unstable intermediate along the pathway and a lower barrier to folding. Pathway two has a much higher barrier to folding, accounting for its rarer observation in folding trajectories. Since $T = 0.43$ is below the folding temperature, we see that in both cases the final folded state is much lower in free energy. The observed kinetics of folding to the native state at this temperature will result from trajectories following paths similar to pathway one or pathway two and the ensemble of allowable paths in between these extremes.

Kinetics

We can characterize the rate of the folding process by tabulating the distribution of mean-first passage times at various temperatures. Figure 7 shows this for a range of temperatures, above and below the folding temperature.

At higher temperatures, the kinetics are best fit by a single exponential, indicating a single barrier in the folding process. The folding times reach a minimum near $T = 0.55$ where the competition between faster folding because of the lower barrier to the native state is balanced by the slower folding of collapsed misfolded states. As the temperature is lowered further, we see a crossover to bi-exponential kinetics characterized by a fast phase and a much slower phase. In this regime, some chains fold to the native state immediately upon collapse constituting the fast folding phase. The slower folding phase consists of chains that have initially collapsed to misfolded states and must partially or completely unfold to reach the native state.

However, even at the relatively low temperature of $T = 0.35$, we still see an excellent fit to bi-exponential kinetics. This indicates that we have not entered the glasslike regime where kinetics might be expected more to follow a power law (Nymeyer *et al.*, 1998) or stretched exponential behavior (Phillips, 1995). This establishes that the kinetic glass temperature T_g is well below the folding temperature T_f for this model ($T_f/T_g > 1.3$), the signature of a good folder (Socci and Onuchic, 1994; Nymeyer *et al.*, 1998). This is a significant improvement over the previous all-β model, which encountered glasslike kinetics well before the folding temperature (Nymeyer *et al.*, 1998; Sorenson and Head-Gordon, 1999).

The question of how quickly folding accompanies chain collapse can be addressed by examining the kinetics of chain collapse at different temperatures versus folding to the native state. Even at the fastest folding temperature, chain collapse still occurs an order of magnitude faster than folding (Sorenson and

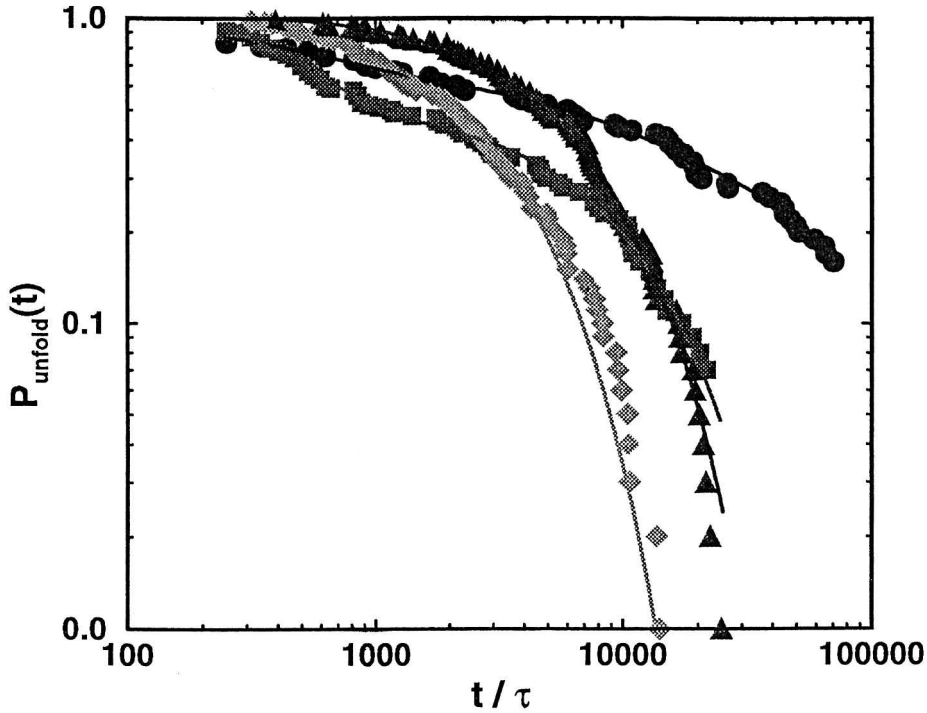

FIG. 7. Percentage of unfolded states versus time at various temperatures for the mixed α/β model: $T = 0.6$ (triangles), $T = 0.55$ (diamonds), $T = 0.45$ (squares), $T = 0.35$ (circles). The solid lines are bi-exponential fits for $T = 0.35$ and $T = 0.45$ and single exponential fits for $T = 0.55$ and $T = 0.6$.

Head-Gordon, unpublished). While this might argue that collapse is not concomitant with folding, this should be contrasted to the scenarios described by other models such as two-flavor lattice models which have collapse and folding times differing by three or more orders of magnitude (Socci and Onuchic, 1995; Sorenson and Head-Gordon, 1998). At the lowest temperature examined, $T = 0.35$, chain collapse occurs two orders of magnitude faster than folding—consistent with the observation above that collapse to misfolded states plays an important role in the kinetics at this temperature.

DISCUSSION

Unlike previous work on similar off-lattice bead models, our model shows an underlying folding funnel that strongly directs the unfolded chain to the folded state. As comparison of Figures 2–4 shows, folding to the native state is closely associated with collapse of the chain. Our model seems to support a collapse-accompanying-folding scenario for this sequence and structure. This is further supported by the observation that the collapse temperatures and folding temperatures are nearly coincident and the observation that the collapse and folding times differ by only an order of magnitude at the fastest folding temperatures. In contrast, earlier studies on the all-β model suggest a collapse-and-then-fold scenario with collapse occurring at significantly higher temperatures than the folding temperature (Sorenson and Head-Gordon, 1999). These observations on the model parallel experimental work on protein L; recent time-resolved small-angle X-ray scattering experiments on protein L have determined that chain collapse is the rate-limiting step, occuring relatively closely in time to native state formation (Plaxco *et al.*, 1999).

Another interesting feature of this model and our analysis is the identification of two distinct pathways for folding to the native state. The existence of parallel folding pathways has been a topic of much current experimental research (Laurents and Baldwin, 1998; Dobson *et al.*, 1998), having been suggested by theoretical research on folding funnels (Dill and Chan, 1997; Onuchic *et al.*, 1997). By using multiple order parameters to characterize the free energy landscape, we can confidently assert the existence of at least two pathways; simply projecting the free energy landscape onto a single order parameter does not uncover these parallel pathways (Sorenson and Head-Gordon, unpublished).

These two pathways are further notable in that one involves a slightly stable intermediate and the other appears to have no stable intermediate. This is important in the current context because experiments on protein G have established the existence of a high-energy intermediate along the folding pathway (Park *et al*., 1997). On the other hand, the folding of protein L, with a nearly identical tertiary structure, appears to be purely two-state with a single barrier (Scalley *et al*., 1997; Yi and Baker, 1996; Plaxco *et al*., 1999). Our model encompasses both folding scenarios and shows how the same structure could form from either pathway. This prompts the experimental question of what parts of the protein G sequence must be changed to modify its kinetics to more pure two-state behavior, or, alternatively, what changes to protein L would induce three-state folding.

The role of the individual secondary structural units in the folding pathway for this model is also comparable to experiment. While the helix is mostly formed in the dominant pathway involving formation of β-hairpin #1, the stability of the helix is weak even at temperatures far below the folding temperature, as can be seen in Figure 3. Also seen in that figure is that the folding transition for the helix is significantly less cooperative than the folding transition for β-hairpin formation. This is consistent with protein engineering experiments on protein L which have demonstrated that the helix plays a spectator role in the folding process and is mostly disrupted at the transition state (Kim *et al*., 1998b). Combined, these results indicate that formation of a well-formed helix is not a necessary condition for folding of this protein to proceed. The asymmetrical formation of the β-hairpin structures seen in Figure 5 is, interestingly, also an observed aspect of the folding of protein L (Gu *et al*., 1997), providing further indication that mutation studies on this model can be fruitfully compared to corresponding experimental studies.

A recurring problem in designing minimalist protein folding models is the encountering of glasslike kinetics at temperatures above the folding temperature (Socci and Onuchic, 1994; Sorenson and Head-Gordon, 1998; Nymeyer *et al*., 1998; Sorenson and Head-Gordon, 1999). Our model appears to avoid this problem, with bi-exponential folding kinetics at temperatures well below the folding temperature. Characterization of our model as a good folder is consistent with the collapse-accompanying-folding scenario described above, since we expect the coincidence of T_θ with T_f to be a signature of good folding models and real proteins (Klimov and Thirumalai, 1996; Klimov and Thirumalai, 1998).

CONCLUSION

In this paper we have introduced a new off-lattice bead model capable of simulating a wide range of small protein structures. This model possesses the advantage of being computationally feasible while at the same time possessing sufficient complexity to allow meaningful comparison with experiment. The model is formulated in such a way that we can investigate various experimentally relevant questions, such as the role of secondary structure versus tertiary structure formation in forming the native state, the effect of changing secondary structure propensities in turn regions or regions of well-defined secondary structure, and the effect of sequence mutations such as destabilizing core mutations.

As an example of the model, we designed a mixed α/β structure with a native state resembling the IgG-binding proteins L and G. By characterizing the thermodynamics and kinetics of the folding of this model, we have shown that the folding process is highly cooperative and the kinetics of folding remain nonglasslike down to temperatures much below the folding temperature, indicating a strong folding funnel. Two distinct folding pathways have been shown to exist, one with a slightly stable intermediate and one without. This is especially relevant in comparison to proteins L and G, since protein L appears to follow purely two-state folding (Scalley *et al*., 1997; Yi and Baker, 1996; Plaxco *et al*., 1999), while the folding of protein G appears to contain a partly stable intermediate (Park *et al*., 1997). From the free energy landscape of our model, we can suggest that the protein G intermediate is nonobligatory.

The mixed α/β model presented here, while computationally simple, shows complex folding behavior that can be usefully compared with experiments on real proteins. In future work, this model will investigate experimental studies where sequence mutations have been introduced into the turn (Gu *et al*., 1997) and helix (Kim *et al*., 1998b) regions. Our model is also able to simulate all-β and all-α proteins. Analyzing the folding of these different fold classes within the same framework should provide insight into the role of local interactions versus nonlocal interactions for the formation of these different fold classes. Lastly, we expect to be able to add simple descriptions of solvent to this new model for small proteins, to better clarify the role of solvation forces in guiding the unfolded protein to fold quickly and correctly to the native state (Sorenson and Head-Gordon, 1998; Sorenson *et al*., 1999).

ACKNOWLEDGMENTS

T.H.G. would like to acknowledge financial support from Air Force Office of Sponsored Research Grant #FQ8671-9601129 and United States Department of Energy Contract #DEAC-03-76SF00098. J.M.S. thanks the National Science Foundation for a graduate research fellowship.

REFERENCES

Blanco, F., Ortiz, A., and Serrano, L. 1997. Role of a nonnative interaction in the folding of the protein G B1 domain as inferred from the conformational analysis of the α-helix fragment. *Fold. Des.* 2, 123–133.

Blanco, F., and Serrano, L. 1995. Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* 230, 634–649.

Capaldi, A., and Radford, S. 1998. Kinetic studies of β-sheet protein folding. *Curr. Opin. Struct. Biol.* 8, 86–92.

Chou, J., and Shakhnovich, E. 1999. A study on local–global cooperativity in protein collapse. *J. Phys. Chem.* B 103, 2535–2542.

Dill, K., Bromberg, S., Yue, K., Fiebig, K., Yee, D., Thomas, P., and Chan, H. 1995. Principles of protein folding—a perspective from simple exact models. *Protein Sci.* 4, 561–602.

Dill, K., and Chan, H. 1997. From Levinthal to pathways and funnels. *Nature Struct. Biol.* 4, 10–19.

Dobson, C., Šali, A., and Karplus, M. 1998. Protein folding: A perspective from theory and experiment. *Angew. Chem. Int. Ed. Engl.* 37, 868–893.

Duan, Y., and Kollman, P. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282, 740–744.

Eaton, W., Muñoz, V., Thompson, P., Henry, E., and Hofrichter, J. 1998. Kinetics and dynamics of loops, α-helices, β-hairpins, and fast folding proteins. *Acct. Chem. Res.* 31, 745–753.

Ferguson, D., and Garrett, D. 1999. Simulated annealing-optimal histogram methods. *Adv. Chem. Phys.* 105, 311–336.

Ferrenberg, A., and Swendsen, R. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63, 1195–1198.

Fersht, A. 1997. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7, 3–9.

Gu, H., Kim, D., and Baker, D. 1997. Contrasting roles for symmetrically disposed β-turns in the folding of a small protein. *J. Mol. Biol.* 274, 588–596.

Guo, Z., and Brooks, C. 1997. Thermodynamics of protein folding: A statistical mechanical study of a small all-β protein. *Biopolymers* 42, 745–757.

Guo, Z., and Thirumalai, D. 1994. Kinetics of protein folding: Nucleation mechanism, time scales, and pathways. *Biopolymers* 36, 83–102.

Guo, Z., and Thirumalai, D., 1996. Kinetics and thermodynamics of folding of a *de novo* designed four-helix bundle protein. *J. Mol. Biol.* 263, 323–343.

Honeycutt, J., and Thirumalai, D. 1990. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci.* 87, 3526–3529.

Kim, D., Gu, H., and Baker, D. 1998a. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci.* 95, 4982–4986.

Kim, D., Yi, Q., Gladwin, S., Goldberg, J., and Baker, D. 1998b. The single helix in protein L is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* 284, 807–815.

Klimov, D., and Thirumalai, D. 1996. Criterion that determines the foldability of proteins. *Phys. Rev. Lett.* 76, 4070–4073.

Klimov, D., and Thirumalai, D. 1998. Cooperativity in protein folding: From lattice models with sidechains to real proteins. *Fold. Des.* 3, 127–139.

Kolinski, A., Galazka, W., and Skolnich, J. 1996. On the origin of the cooperativity of protein folding: Implications from model simulations. *Proteins: Struct. Funct. Genet.* 26, 271–287.

Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R., and Kollman, P. 1995. Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comp. Chem.* 16, 1339–1350.

Laurents, D., and Baldwin, R. 1998. Protein folding: Matching theory and experiment. *Biophys. J.* 75, 428–434.

Liwo, A., Kaźmierkiewicz, R., Czaplewski, C., Groth, M., Ołdziej, S., Wawak, R., Rackovsky, S., Pincus, M., and Scheraga, H. 1998. United-residue force field for off-lattice protein-structure simulations: lii. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J. Comp. Chem.* 19, 259–276.

Muñoz, V., Thompson, P., Hofrichter, J., and Eaton, W. 1997. Folding dynamics and mechanism of β-hairpin formation. *Nature* 390, 196–199.

Nymeyer, H., García, A., and Onuchic, J. 1998. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci.* 95, 5921–5928.

Onuchic, J., Luthey-Schulten, Z., and Wolynes, P. 1997. Theory of protein folding: The energy landscape perspective. *Ann. Rev. Phys. Chem.* 48, 545–600.

Park, S.-H., O'Neil, K., and Roder, H. 1997. An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry* 36, 14277–14283.

Philips, J. 1995. Kohlrausch relaxation and glass transitions in experiment and in molecular dynamics simulations. *J. Non-Cryst. Solids* 182, 155–161.

Plaxco, K., Millett, I., Segel, D., Doniach, S., and Baker, D. 1999. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nature Struct. Biol*. 6, 554–556.

Ramírez-Alvarado, M., Serrano, L., and Blanco, F. 1997. Conformational analysis of peptides corresponding to all the secondary structure elements of protein L B1 domain: Secondary structure propensities are not conserved in proteins with the same fold. *Protein Sci*. 6, 162–174.

Scalley, M., Yi, Q., Gu, H., McCormack, A., Yates, J., and Baker, D. 1997. Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry* 36, 3373–3382.

Shakhnovich, E. 1994. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett*. 72, 3907–3910.

Shakhnovich, E. 1996. Modeling protein folding: The beauty and power of simplicity. *Fold. Des*. 1, R50–R54.

Shakhnovich, E. 1997. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol*. 7, 29–40.

Shea, J.-E., Nochomovitz, Y., Guo, Z., and Brooks, C. 1998. Exploring the space of protein folding Hamiltonians: The balance of forces in a minimalist β-barrel model. *J. Chem. Phys*. 109, 2895–2903.

Socci, N., and Onuchic, J. 1994. Folding kinetics of proteinlike heteropolymers. *J. Chem. Phys*. 101, 1519–1528.

Socci, N., and Onuchic, J. 1995. Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. *J. Chem. Phys*. 103, 4732–4744.

Sorenson, J., and Head-Gordon, T. 1998. The importance of hydration for the kinetics and thermodynamics of protein folding: Simplified lattice models. *Fold. Des*. 3, 532–534.

Sorenson, J., and Head-Gordon, T. 1999. Redesigning the hydrophobic core of a model β-sheet protein: Destabilizing traps through a threading approach. *Proteins: Struct. Funct. Genet*. 37, 582–591.

Sorenson, J., and Head-Gordon, T. unpublished.

Sorenson, J., Hura, G., Soper, A., Pertsemlidis, A., and Head-Gordon, T. 1999. Determining the role of hydration forces in protein folding. *J. Phys. Chem*. B 103, 5413–5426.

Wikstrom, M., Drakenberg, T., Forsen, S., Sjobring, U., and Bjoerck, L. 1994. Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L—comparison with the IgG-binding domain of protein G. *Biochemistry* 33, 14011–14017.

Wolynes, P. 1997. As simple as can be? *Nature Struct. Biol*. 4, 871–874.

Yi, Q., and Baker, D. 1996. Direct evidence for a two-state protein unfolding transition from hydrogen–deuterium exchange, mass spectrometry, and NMR. *Protein Sci*. 5, 1060–1066.

Yue, K., Fiebig, K., Thomas, P., Chan, H., Shakhnovich, E., and Dill, K. 1995. A test of lattice protein folding algorithms. *Proc. Natl. Acad. Sci*. 92, 325–329.

Address correspondence to:
Teresa Head-Gordon
Physical Biosciences and Life Sciences Divisions
Lawrence Berkeley National Laboratory
One Cyclotron Road
Donner 472
Berkeley, CA 94720